Scaling the Critics: Uncovering the Latent Dimensions of Movie Criticism with an Item Response Approach

Scaling the Critics

Uncovering the Latent Dimensions of Movie Criticism with An Item Response Approach*

Michael Peress† Arthur Spirling‡

June 1, 2009

*Excellent research assistance from Edward Laird and Chris Tice is gratefully acknowledged. We thank Brett Gordon and Keith Poole for useful comments. This work was originally presented as a poster at the Summer Political Methodology Meeting (2008) and we thank participants for feedback, especially Chris Achen and Alastair Smith. Peress thanks the Institute of Quantitative Social Science for hospitality. We are very grateful for comments from two anonymous referees and the AE at *JASA* that helped us improve the content and structure of our paper.

†Department of Political Science, University of Rochester. mperess@mail.rochester.edu

‡Department of Government and Institute of Quantitative Social Science, Harvard University. aspirling@gov.harvard.edu

Abstract

We study the critical opinions of expert movie reviewers as an item response problem. Building on earlier 'unfolding' models, we develop a framework that models an individual's decision to approve or disapprove of an item. Using this approach, we are able to recover the locations of movies and ideal points of critics in the same multi-dimensional space. We demonstrate that a three dimensional model captures much of the variation in critical opinions. The first dimension signifies movie 'quality' while the other two connote the nature and subject matter of the films. We then demonstrate that the dimensions uncovered from our 'utility threshold model' are statistically significant predictors of a movie's success, and are particularly useful in predicting the success of 'independent' films.

Key words: *utility threshold model* *film* *ideal points*

1 Introduction

For the year 2006, the Motion Picture Association reported that international revenues generated by its composite companies totaled some $42.6 billion (Hollinger, 2007). This sum is on a par with the gross domestic product of Kenya for the same period. Clearly then, the movie industry is an important economic force both in the United States ($24.3 billion revenue for 2006) and elsewhere ($18.3 billion). Fulfilling a consumer-advisory rôle within this massive sector, movie *critics* are ubiquitous: reviews and recommendations for films can be found in many journalistic outlets like newspapers, magazines and online websites. Major studios apparently accord substantial influence to such critics, as do film historians: Smith (1998), for example, names the critics Gene Siskel and Roger Ebert in his top 100 ranking of the most influential people in movie history. Critics are fêted with press kits, advance screenings and other perks, and then using (selected, positive) reviewers' opinions directly in the marketing of their product. Indeed, Sony Pictures went so far as to create a fictional critic—named David Manning—whose enthusiastic (and entirely fabricated) 'quotes' appeared on several of the studio's movie adverts circa 2001 (Elsworthin, 2005).

Quite apart from their significance to large film-making firms and the news media devoted to the entertainment-industry, there is considerable *academic* interest in critics' choices and decision-making processes. First, within the marketing literature, assessing and quantifying the influence of critical reception on the commercial success of film media has been an ongoing concern (Ainslie, Drèze and Zufryden, 2005; Eliashberg and Shugan, 1997; Neelamegham and Chintagunta, 1999). Modeling the behavior of critics directly would thus paint a more complete picture of the interrelationship between film characteristics and market performance. Second, film criticism—particularly when practiced by those versed in *film theory* —is an important element of cultural studies, a discipline that seeks to systematically understand cultural phenomena in terms of their social, political and psychological causes

and consequences. Hence, there is motive to explore the ways in which audiences 'receive' the motion picture medium (Blumer, 1933; Kracauer, 1957; Mulvey, 1975; Riesman, Denny and Glazer, 1968). By analyzing new data on hundreds of critical reviews this paper seeks to contribute to both these scientific endeavors.

As we will describe in more detail below, the data are examples of item responses (Lord, 1980; Hambleton, Swaminathan and Rogers, 1991). Our central concern is using psychometric measurement techniques—especially those derived from item response theory (IRT)—to uncover the latent traits that characterize movie critics and the movies that they review. The data differ from traditional applications since the subjects here choose whether to 'approve' or 'disapprove' of a *single* item. Hence, our theoretical framework of actor behavior leads us to employ a statistical approach that differs somewhat from the cumulative models commonly seen in social science applications like educational testing (Rasch, 1961; Lord, 1980; Bock and Aitken, 1981), marketing research (Kamakura and Srivastava, 1986; Goettler and Shachar, 2001; Anand and Byzalov, 2008), and legislator ideal point estimation (Poole and Rosenthal, 1997; Martin and Quinn, 2001; Clinton, Jackman and Rivers, 2004). Specifically when uncovering legislative ideal points, notice that the spatial locations of *two* alternatives are of interest: the status quo and the proposal; this is a sharp contrast to the critic case, where only *one* alternative is reviewed. Our framework—the 'utility threshold model'—applies to movie criticism, and more generally, to approval or ordinal rating data.

Our paper generalizes existing models for approval data in a number of ways. First, our framework is multidimensional. This is important because we expect critics to differ in their preferred movie characteristics. Second, we allow for a non-diagonal proximity metric in our estimation. As we show in the paper, this is necessary for preserving rotational invariance in the model. Third, we allow critics to differ in their approval thresholds. This feature is necessary to account for the fact that some critics are stingier with their praise than others. Fourth, we can recover the ideal points of critics and the locations of movies in

the same multidimensional space. This differentiates our procedure from scaling procedures developed for dichotomous and polytomous choice data. Finally, when applied to ratings data, our procedure allows us to control for a type of selection bias which may be present in indices of movie quality. Specifically, critics may choose to review movies that they expect to enjoy. Our procedure can control for this type of selection bias, if the critics' choices of which movies to review are based on the spatial characteristics of the movies.

Intriguingly, we find that the 'expert' critics in our data set—and the movies themselves—are almost fully described by three latent dimensions: they pertain to 'quality', followed by a division of space between 'nerds', 'jocks' and 'art-house'. These latter labels refer to types of consumers who might enjoy predominantly science-fiction, action adventure and deep (potentially disturbing) emotional movies, respectively. We demonstrate that such reviews are good predictors of financial success for movie makers, especially for independent films with relatively narrow audiences.

Outside of movie criticism, our estimator applies to a number of other important problems. Legislators choose whether or not to cosponsor legislation. In marketing, a panel of consumers may be given a set of products to rate, and latent characteristics of these products could be deduced from these ratings. In admissions processes to universities, officers decide whether or not to allow a potential students entry based on their qualities. More generally, our framework extends existing models for approval data in a way necessary for analyzing the diversity of choice present in many applications.

2 Data and Background

2.1 Data

Until relatively recently, data on critic responses to movies was both widely-scattered and in no standard form: different media recorded reviews in multiple ways, from long discursive

articles with implicit judgments, to spoken television or radio reports to summary 'star'-system recommendations. It was thus extremely costly to collate critical opinions. Moreover, the analyst was typically required to either use a few 'key' reviewers as indicative of a larger audience, or laboriously recode responses in order to make them comparable. The advent of the internet, however, has changed matters. *Rotten Tomatoes*, a website situated at `http://www.rottentomatoes.com`, collates both multiple reviews for any given movie, and codes each review—in terms of how positive or negative it was towards the film—using a common rating system. In particular, *Rotten Tomatoes* considers each film review by each different critic (of which more than 100 may exist for recent movies) and then denotes the opinion as 'fresh' (i.e. the critic recommends the film) or 'rotten' (i.e. the critic does not recommend the film). This information is available to the public.

To see how this information might be used, first let $c = 1, \ldots, C$ index the critics and let $m = 1, \ldots, M$ index the movies. The data to be modeled is then a $C \times M$ matrix of observed ratings (coded by *Rotten Tomatoes*) by the C critics on the M movies. Let $\mathbf{Y}$ denote this matrix and let $y_{c,m}$ denote the rating critic c gives movie m. We will code $y_{c,m} = 2$ if the critic recommended the movie (i.e. it is 'fresh'), $y_{c,m} = 1$ if the critic did *not* recommend it (i.e. it is 'rotten'), and $y_{c,m} = 0$ if the critic did not review the movie.

Our database uses a very expansive definition of what it is to be a film critic. Individuals who submit only a handful of film reviews to online mailing lists are considered critics. To focus on the population of interest—expert reviewers—we restrict $\mathbf{Y}$ to all critics who are members of the *National Society of Film Critics*. This organization holds a prestigious place within the movie reviewing world and consists of approximately 60 respected individuals, all of whom are elected to their positions. In addition, these critics typically write and turn in their reports for publication at approximately the same time. Thus there is little danger, for example, that critics respond to *each other's opinions* rather than their viewing experience. We included all such critics who reviewed at least 20 films and all films that received at least

50 reviews on *Rotten Tomatoes*. The resulting dataset has approximately 50 critics and 1000 movies. The minimum number of reviews a movie received among the NSFC critics was 16, while the median number of movies each NSFC critic reviewed was 336.

2.2 Cumulative Models

As should be clear, the matrix $\mathbf{Y}$ contains rows of 'individuals' responding in a dichotomous way to 'items' in its columns. If we wish to understand the latent traits possessed by both critics and movies, IRT seems a reasonable way to proceed. It is quite common to consider the following model,

$$\Pr(y_{c,m} = 2) = F(a_m(\theta_c - b_m)). \tag{1}$$

where F represents a strictly increasing cumulative distribution function (cdf). When F is chosen to be the Gaussian cdf, we have the normal ogive model. When F is chosen to be the logistic distribution (i.e. $F(x) = 1/(1 - e^{-x})$), we have Birnbaum's two-parameter logistic model. When F is chosen to be the logistic distribution and $a_m = 1$ for all m, we have the Rasch model as a special case.

These approaches are collectively referred to as *cumulative* models. When applied to educational testing, θ_c is interpreted as the intelligence of individual c, b_m is interpreted as the difficulty of item m, and a_m determines the discrimination power of item m. Variants of these models allow for more than two responses, multiple dimensions of intelligence, a nonzero probability of guessing a correct answer, and various other features. Such models share the property that the probability of observing a 'correct' response of $y_{c,m} = 2$ is strictly increasing in intelligence θ_c. This is reasonable when applied to education testing, but may not be appropriate in some other applications.

2.3 Unfolding Models

An alternative to the cumulative model is the *unfolding* model, pioneered by Coombs (1964, esp. Ch 15). The unfolding model differs from the cumulative model in that the probability of a positive response is strictly decreasing in the distance between an individual's ideal point and the spatial location of the item. The probability of observing a positive response is maximized at the individual's ideal point, denoted by α_c.

It is this framework that we build upon in our model of movie criticism. The unfolding model often takes the form,

$$\Pr(y_{c,m} = 2) = F(-(\alpha_c - \delta_m)^2). \tag{2}$$

Here, F would typically be selected to the logistic or normal distribution. Examples of unfolding models include DeSarbo and Hoffman (1987), Andrich (1988), Hoijtink (1990; 1991), Andrich and Luo (1993), Takane (1996), Leenen and Mechelen (2004) and Maydeu-Olivares, Hernandez and McDonald (2006). These models differ in the exact set of assumptions they employ, including whether the characteristic space is allowed to be multidimensional, whether the ideal points are treated as fixed or random effects, and so on.

3 Model and Estimation Procedure

3.1 The Utility Threshold Model

Our model should have a number of features. First, it should be multidimensional because we expect critics to differ in their preferred movie characteristics. Second, the model should be of an unfolding variety. This will allow critics to prefer movies that offer a combination of some action and some romance, for example. Cumulative models, by contrast, would require critics to have preferences that are strictly increasing (or decreasing) in 'action-ness'.

Third, we should allow for a non-diagonal weighting matrix. This is mostly a technical requirement, but is necessary to ensure that the resulting likelihood function is invariant to linear transformations of the characteristic space. A fourth requirement is that critics with similar ideal points should be allowed to differ in the probability that they assign a given movie a positive review. Some critics may simply be 'stingier' with their praise, and we would like to be able to capture this in our framework.

We begin by assuming that the ideal points of critics and the locations of movies can be represented in the same D-dimensional space. We let $\boldsymbol{\alpha}_c \in \mathbb{R}^D$ denote the ideal point of critic c and we let $\boldsymbol{\delta}_m \in \mathbb{R}^D$ denote the location of movie m. For example, there might be three dimensions (i.e. $D = 3$) in which all movies and critics can be situated: perhaps the first dimension corresponds to 'action-ness', the second to 'romance-ness' and the third to 'drama-ness'. A romantic-comedy would have a 'low' score on the first dimension, but be 'high' on the other two. It seems sensible to suppose that critics are most likely to approve of a movie that is close to their ideal point, and we assume the utility critic c gets from movie m is given by,

$$u_{c,m} = -(\boldsymbol{\alpha}_c - \boldsymbol{\delta}_m)'\boldsymbol{W}(\boldsymbol{\alpha}_c - \boldsymbol{\delta}_m) + \epsilon_{c,m}. \tag{3}$$

Here, $\epsilon_{c,m}$ are independent and identically distributed shocks from the standard normal distribution, and $\boldsymbol{W}$ is a symmetric positive definite weighting matrix. A critic who likes romantic-comedies over all other types of films would have an ideal point which is low on the first dimension and high on the other two. We assume that critic c gives a positive review to movie m if his utility exceeds his approval threshold. Hence, we observe a fresh rating if $u_{c,m} \geq \bar{u}_c$, or equivalently,

$$\epsilon_{c,m} \geq \bar{u}_c + (\boldsymbol{\alpha}_c - \boldsymbol{\delta}_m)'\boldsymbol{W}(\boldsymbol{\alpha}_c - \boldsymbol{\delta}_m). \tag{4}$$

One may worry that critics choose to review movies that they expect to like (because they

enjoy seeing good movies) or movies that they expect to dislike (to allow for entertaining reviews). These facts are accounted for in our framework, to the extent that such selection operates on the estimated critic ideal points and movie locations. This is true because we explicitly model the process by which critics decide whether to like or dislike a movie in terms of movie locations and critic ideal points. In this sense, we control for many of the aspects that determine whether a critic likes or dislikes a movie.

Returning to the derivation, we have,

$$\Pr(y_{c,m} = 1) = \Phi(\bar{u}_c + (\boldsymbol{\alpha}_c - \boldsymbol{\delta}_m)'\boldsymbol{W}(\boldsymbol{\alpha}_c - \boldsymbol{\delta}_m)) \quad (5)$$

$$\Pr(y_{c,m} = 2) = 1 - \Phi(\bar{u}_c + (\boldsymbol{\alpha}_c - \boldsymbol{\delta}_m)'\boldsymbol{W}(\boldsymbol{\alpha}_c - \boldsymbol{\delta}_m)). \quad (6)$$

Note that the 'zero'-dimensional model is of interest as well. In this case, there are no *spatial* locations of either critics or movies to be estimated: movies are treated as homogenous entities and the only source of heterogeneity comes from the fact that some critics are stingier with their praise.

As a way to interpret the model in Equations (5) and (6), consider the 'trace line' in Figure 1. Notice that for any particular utility threshold, $\bar{u}$, the critic's probability of approving the film is decreasing quadratically as the movie's location ($\bar{\delta}$) moves away from his spatial preference (α)—he is most likely to approve when their locations coincide (when $\alpha - \bar{\delta} = 0$). For any particular spatial distance between movie and critic, notice that increasing $\bar{u}$ (i.e. making the critic generally harder to please) will decrease the probability that he approves of the movie.

[Figure 1 about here.]

We can write the log-likelihood function as follows,

$$\mathcal{L}^{C,M}(\boldsymbol{\alpha}, \bar{u}, \boldsymbol{\delta}, \boldsymbol{W}) = \sum_{c=1}^{C}\sum_{m=1}^{M}\left[1\{y_{c,m}=1\}\log\Phi\left(\bar{u}_c + (\boldsymbol{\alpha}_c - \boldsymbol{\delta}_m)'\boldsymbol{W}(\boldsymbol{\alpha}_c - \boldsymbol{\delta}_m)\right)\right.$$

$$\left. +1\{y_{c,m}=2\}\log\left\{1-\Phi\left(\bar{u}_c + (\boldsymbol{\alpha}_c - \boldsymbol{\delta}_m)'\boldsymbol{W}(\boldsymbol{\alpha}_c - \boldsymbol{\delta}_m)\right)\right\}\right]. \quad (7)$$

Estimating the parameters of the model can be accomplished by maximizing (7). This is straightforward in principle, but a number of complication arise, which we describe later in this section.

3.2 Relationship to Applied IRT

As noted above, one of the simplest educational models (see Lord, 1980, for example) has a test taker with latent trait z'_m determining her performance on item m. The trait z'_m is a composite of the examinee's ability θ and an error component for item m. Typically, we assume those errors are normally distributed, and that they have equal variance regardless of the ability of the students concerned. Before going further, notice that for our model, that assumption about disturbances yields the representation in Figure 2. Here, there are three individuals h, j, k with different spatial preferences (α parameters) but the same utility threshold $\bar{u}$. They are confronted with the same movie which we will assume has $\delta = 3$. Recall that utility is quadratically decreasing in the movie's distance from the reviewer. Critic h has a spatial preference for $\alpha \approx 1$ so he is likely to disapprove of the movie. By contrast, j is most likely to approve—and more likely to do so than k. Now suppose that we abandon the assumption of a common $\bar{u}$, such that k is more difficult to please (though her spatial preferences are similar to before). The shift up from $\bar{u}$ to $\bar{u}_k$ will make k more likely to disapprove of the movie: a larger portion of her error term is now shaded.

[Figure 2 about here.]

The estimator we propose is closely related, but is it not isomorphic, to the estimators commonly used for item response theory. Our estimator can be written as,

$$\Pr(y_{c,m} = 2) = F(\bar{u}_c + \boldsymbol{\alpha}_c{}'\boldsymbol{W}\boldsymbol{\alpha}_c - 2\boldsymbol{\alpha}_c{}'\boldsymbol{W}\boldsymbol{\delta}_m + \boldsymbol{\delta}_m{}'\boldsymbol{W}\boldsymbol{\delta}_m) \tag{8}$$

while the multidimensional normal ogive model can be written as,

$$\Pr(y_{c,m} = 2) = F(\boldsymbol{a}_m + \boldsymbol{b}_m'\boldsymbol{\theta}_c) \tag{9}$$

We can set up a relationship between the two models by letting $\boldsymbol{\theta}_c = ([\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{\alpha}_c]_1, \ldots, [\boldsymbol{W}^{\frac{1}{2}}\alpha_c]_D, \bar{u}_c + \boldsymbol{\alpha}_c{}'\boldsymbol{W}\boldsymbol{\alpha}_c)$, $\boldsymbol{a}_m = \boldsymbol{\delta}_m{}'\boldsymbol{W}\boldsymbol{\delta}_m$, and $\boldsymbol{b}_m = (-2[\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{\delta}_m]_1, \ldots, -2[\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{\delta}_m]_D, 1)$. We now have that the D-dimensional utility threshold model is isomorphic to a $D+1$-dimensional normal ogive model where the last component of $\boldsymbol{b}_m$ is restricted to be equal to 1. Otherwise put, we can always find a $D+1$-dimensional normal ogive model which summarizes the data at least as well as the D-dimensional utility threshold model, and we can always find a $D+1$ utility threshold model which summarizes the data at least as well as a D-dimensional normal ogive model. This arrangement suggests that we cannot differentiate between the utility threshold and normal ogive models on the basis of model-fit alone (and hence, we don't try to).

Instead then, the advantage of the utility threshold model is that it posits an appropriate structural model for the data, which allows us to correctly interpret the estimated parameters when applied to movie criticism data (and approval or ordinal rating data more generally). If the true data generating process were a D-dimensional utility threshold model, we would be able to successfully fit a $D+1$-dimensional normal ogive model. The difficulty would come in interpreting $\boldsymbol{\theta}_c$ and $(\boldsymbol{a}_m, \boldsymbol{b}_m)$. Note that $\boldsymbol{\theta}_c$ would contain the same information as $(\alpha_{c,1}, \ldots, \alpha_{c,D}, \bar{u}_c)$, but the estimates would not reveal which components of $\boldsymbol{\theta}_c$ characterize

the ideal points and which components characterize heterogeneity in the thresholds. This problem occurs because of the rotational invariance present in item response models, meaning that $\bar{u}_c$ need not appear as the last element of $\boldsymbol{\theta}_c$.

A second advantage of our technique is that we can recover critic and movie locations in the same multidimensional space, something which would be impossible if we applied the traditional item response estimator to approval data. Cumulative models are closely related to the dichotomous choice models considered in the political science, economics, and marketing literatures. In these dichotomous choice models, individuals choose between two items located in a multi-dimensional space. Each individual has an ideal point located in the same-multidimensional space. This framework has a reduced form that is isomorphic to the multi-dimensional cumulative model. In applications of the normal ogive model to dichotomous choice data, we can recover ideal points and cutting planes in the same multidimensional space, but we cannot recover item *locations* because we cannot separately identify the distance between the items and the variance of the disturbance term for that item (Poole, 2005). Our setup is different because individuals rate a single item at a time. This is the key difference that allows us to recover movie locations in our framework.

3.3 Identification

As is usual with such models we must impose some restrictions on the parameters in order to ensure identification. In the case of the standard multi-dimensional item response problem, it is well known that $\boldsymbol{\theta}_c$ must be constrained for $D+1$ individuals. A similar solution emerges here.

Throughout, we use zero subscripts to denote the parameters of the data generating process, i.e. the "true" parameter values. That is, $\boldsymbol{\alpha}_0 = (\alpha_{1,0}, \ldots, \alpha_{C,0})$ denote the true critic ideal points, $\boldsymbol{\delta}_0 = (\delta_{1,0}, \ldots, \delta_{M,0})$ denote the true movie characteristics, etc. Unsurprisingly, the parameters of the utility threshold model are only identified up to location and scale.

Specifically, consider the reparametrization,

$$\boldsymbol{\alpha}_c = \boldsymbol{A}\boldsymbol{\alpha}_{c,0} + b, \qquad \bar{u}_c = \bar{u}_{c,0}, \qquad \boldsymbol{\delta}_m = \boldsymbol{A}\boldsymbol{\delta}_{m,0} + b, \qquad \boldsymbol{W} = (\boldsymbol{A}')^{-1}\boldsymbol{W}_0\boldsymbol{A}^{-1} \tag{10}$$

where $\boldsymbol{A}$ has full rank. It is straightforward to show that

$$F(\bar{u}_c + (\boldsymbol{\alpha}_c - \boldsymbol{\delta}_m)'\boldsymbol{W}(\boldsymbol{\alpha}_c - \boldsymbol{\delta}_m)) = F(\bar{u}_{c,0} + (\boldsymbol{\alpha}_{c,0} - \boldsymbol{\delta}_{m,0})'\boldsymbol{W}_0(\boldsymbol{\alpha}_{c,0} - \boldsymbol{\delta}_{m,0}))$$

for all c, m. This indicates that we can apply a linear transformation to the critic ideal points without changing the value of the log-likelihood function, provided we can alter the other parameters in the model. To achieve point identification, we can normalize any $D+1$ ideal points. Without loss of generality, we can constrain $\boldsymbol{\alpha}_{D+1} = \mathbf{0}$ and $\boldsymbol{\alpha}_c = \boldsymbol{e}_c$ for $c \in \{1, \ldots, D\}$ where $\boldsymbol{e}_c$ is a unit vector. These constraints allow us to pin down the location and scale of the critic ideal points and movie locations. Otherwise put, the estimated parameter vector uniquely gives rise to the data seen in practice: there exists no other vector that could possibly be responsible for the data. In the Appendix, we prove that the utility threshold model is identified under these conditions. We effectively show that once we constrain the ideal points of $D+1$ critics, we cannot alter the parameter space leaving the value of the log-likelihood intact, with any transformation (linear or nonlinear).

3.4 Implementation

The utility threshold model bears a strong resemblance to the item response models popular in the psychometric, marketing, and political science literatures. The estimation approaches used fall into three broad categories. Fixed effects estimators treat both the item characteristics and individual characteristics as parameters to estimate (Lord, 1980; Poole and Rosenthal, 1997). Random effects integrate out the item (or individual) characteristics (Bock

and Lieberman, 1970; Bock and Aitken, 1981). Conditional fixed effect estimators concentrate out the item parameters (Rasch, 1961). The fixed effects estimators have the advantage of producing additional information, which in our case includes both the individual (critic) and item (movie) specific parameters. Hence we take this approach. In other applications, we may observe a large number of raters rating a small number of items. In these situations, a random effects model would be more appropriate if the goal is to recover only the item characteristics.

A second choice we must make is whether to employ a maximum likelihood or Bayesian estimator. Both maximum likelihood (Lord, 1980; Poole and Rosenthal, 1997) and Bayesian (Albert, 1992; Beguin and Glas, 2001; Martin and Quinn, 2001) versions of the fixed effects estimator have been applied in the social science literature. Programs for implementing these estimators are widely available but they cannot be *directly* applied here since, as noted, the information we wish to garner is not forthcoming from a standard item-response model. The Bayesian estimator is easier to implement efficiently, and modifying the existing code would not be very difficult. Experience indicates that the maximum likelihood estimator is more difficult to implement, yet it is computationally more efficient, particularly when the dimensionality is large. Because computational efficiency was a chief concern, we choose to implement the latter.

While maximizing the likelihood defined in equation (7) is straightforward in principle, a number of complications arise. First, this model involves a very large number of parameters—$K = C(D+1) + MD + D(D+1)/2$. For example, in a four dimensional model, there are more than 6,000 parameters to estimate. This optimization problem would usually be infeasible, but the special form of the objective function makes it tractable. In particular, we can compute the objective function, the gradient, and the Hessian in $O(CM)$ operations, which is significantly less than the $O(C^2M^2)$ and $O(C^3M^3)$ operations that would usually be required to compute them, respectively. Our implementation relies on the *Zig-Zag* algorithm

that has been applied to estimate nonlinear fixed effects models (Heckman, 1981) and item response models (Lord, 1980; Poole and Rosenthal, 1991, 1997).

A second concern is that despite our restriction to the NSFC critics there is still some sparseness in the data: some movies have few reviews while some critics opine on few films. There is thus potential perfect-separation in the data. For these reasons, we use a penalized-likelihood approach (in the sense of Firth, 1993). Here, we follow the spirit rather than the letter of Firth's suggestions: we do not use a penalization based on Jeffrey's priors and we are not *per se* interested in asymptotic refinements.

That objective function takes the following form:

$$\tilde{\mathcal{L}}^{C,M}(\boldsymbol{\alpha}, \bar{u}, \boldsymbol{\delta}, \boldsymbol{W}) = \mathcal{L}^{C,M}(\boldsymbol{\alpha}, \bar{u}, \boldsymbol{\delta}, \boldsymbol{W}) + \sum_{c=1}^{C} \lambda_u(\bar{u}_c^2) + \sum_{c=1}^{C} \lambda_\alpha(\boldsymbol{\alpha}_c{}'\boldsymbol{\alpha}_c) + \sum_{m=1}^{M} \lambda_\delta(\boldsymbol{\delta}_m{}'\boldsymbol{\delta}_m) \quad (11)$$

where $\mathcal{L}^{C,M}$ is as given in Equation (7) and $\lambda_u > 0$, $\lambda_\alpha > 0$ and $\lambda_\delta > 0$ are penalty terms. An equivalent formulation is to think our approach as finding the mode of the posterior distribution where independent normal priors are placed on $(\boldsymbol{\alpha}, \bar{u}, \boldsymbol{\delta})$ and a degenerate uniform prior is placed on $\boldsymbol{W}$. Notice that the contribution of the penalty terms in the objective function approaches zero as the sample size increases: this is because the likelihood term from Equation (7) involves a double sum while each component of the penalty involves a single term.

4 Results

We estimated a series of models, from zero through eight possible dimensions. Our first task was to choose between these models. We chose not to rely on purely statistical measures of model fit (e.g. a likelihood ratio test) because such measures tend to favor very high-dimensional models in large data sets—far more dimensions than we will be able to

successfully interpret (van der Linden and Hambleton, 1997; Ostini and Nering, 2006). We instead considered the geometric mean probability (the average probability of a correct prediction). Relying solely on in-sample measures of model fit can lead to over-fitting, so we also computed the geometric mean probability just on a holdout sample. In computing the out-of-sample fit, we relied on a 20 percent holdout sample and computed the geometric mean probability among all movies that were reviewed by at least 12 critics. Table 1 displays these measures for the various models.

[Table 1 about here.]

Our choice of dimensionality was based primarily on out of sample fit, but we also considered our ability to interpret the estimated dimensions and the usefulness of the estimated dimensions for subsequent analysis. Using the out of sample geometric mean probability, we found that the three dimensional model was best—it had a geometric mean probability of 64.6%. The baseline model with no spatial dimensions provided a geometric mean probability of 54.3%. Among the models that we estimated, moreover, the dimensions generated by the 3 dimensional model proved easiest to interpret. In addition, we found that the results were most useful for subsequent analysis (such as the regressions we consider in Section 5). Given that these three criteria lead us to the *same* model choice, we are fairly confident that the three dimensional model is most appropriate for this data.

The model we estimated located the movies and critics in three dimensions while also estimating the individual-level utility thresholds for the critics. Recall that a *lower* $\bar{u}$ implies a more permissive critic who *ceteris paribus* is more willing to return a recommendation for the movie. After plotting the density of the thresholds, there is evidence of a slight negative skew: otherwise put, while the majority of critics are symmetrically located, there are a few 'easily pleased' individuals to the far left (see Figure 3). Interestingly, the most generous critic is Roger Ebert (of the *Chicago Sun-Times*) who gives a 'fresh' rating 64% of the time.

It is, by contrast, hard work to impress Amy Taubin, who writes columns for *The Village Voice*—she likes just 39% of the movies she reviews.

[Figure 3 about here.]

In Figure 4 we present a plot of the three spatial dimensions. For the moment, we do not label the points, but they can be demarcated by their shape: the movies appear as round points, while the critics are triangles. A feature of Figure 4 is that the point clouds for critics and movies overlap, but not to the same extent in all dimensions. In the top and middle panels, the movies and critics overlap much less than in the bottom panel. Otherwise put, the δ_1, α_1 dimension appears to discriminate between the groups in space. In particular, the critics generally appear to *right* of the movies: the critics have higher estimated positions on this dimension. To be clear here, under our original normalization, we discovered a dimension with a very high level of discrimination between critic and movie locations. We identified this as a quality dimension and rotated the data (exploiting rotational invariance) such that this dimension appeared as δ_1, to aid in our interpretations.

[Figure 4 about here.]

We contend that this dimension represents a movie's 'quality' and, as we noted earlier, all else equal, critics prefer higher-quality movies to lower-quality ones. In our understanding, 'high quality' movies have a combination of two elements—artistic pretension and production values. Both refer to the craft and ingenuity of movie-making and we would expect 'low quality' movies to include so-called 'B-movies', pornographic and 'exploitation' films. To verify this notion, we conducted the probit regression reported in Table 2. Here, the response is ordered in three categories: 'winner', 'nominated' and 'not nominated' for 'Best Picture' and 'Best Director' at the Academy Awards. The predictor is the movie's estimated δ_1 score, which is significant for both regressions at the $p < 0.01$ level. We obtain similarly significant results when we use the Golden Globe 'Best Motion Picture: Drama' and 'Best Director.'

[Table 2 about here.]

In our conception, for 'expert' critics, quality is associated with the 'high-mindedness' of the movie as art, so small independent films could certainly be included within the rubric. High quality films might well be over-represented in certain genres such as romances, dramas and thrillers rather than, say, horror or action movies. We comment on this below. In Figure 5 we plot the density (and provide a histogram) of both the critics and movie estimates in δ_1, α_1 space—the dimension we claim is quality.

[Figure 5 about here.]

Notice that there is some variance in the estimates for the critics; in our interpretation, this is due to sampling error rather than differing tastes for quality: *ceteris paribus* critics prefer high quality movies, but this does not mean that, say, a higher quality comedy is preferred to a lower quality drama.

Since we are sometimes dealing with relatively small numbers of reviews (e.g. *The Skeleton Key* of 2005 was reviewed by just four NSFC critics), there are reasonably large variances associated with our estimated movie qualities too. To avoid potentially misleading inferences then, in Table 3 we give some ranking information for the films in our sample at the 0.05, 0.5 (i.e. median) and 0.95 quantiles of their empirical cdf of the estimates for δ_1. We also report the `rottentomatoes.com` aggregate ('percent fresh') rating for the movies and, in the final column, the genre description words given for the movies on the site. Notice that our δ_1 dimension estimates seem to agree with the aggregate ratings from the website; moreover, the genres seem fairly uniformly spread throughout the quality distribution, suggesting that this first dimension is indeed quality.

[Table 3 about here.]

From an initial inspection of the movies in the other dimensions δ_2 and δ_3, it was not immediately obvious what these aspects of movie criticism actually were. For example, *The*

Dreamers, a French movie that deals with the sexual awakening of three teenagers during the strife of the 1968 Paris riots seems somewhat different in nature to *Alexander*, a big budget historical epic starring Colin Farrell. Nonetheless these movies inhabit practically the same locations in space. We suspect an explanation lies in the nature of the first, 'quality', dimension of movie review. Put broadly, we would contend that 'bad' movies are actually very similar to one another: a bad comedy is not funny, a bad drama is not very dramatic, and a bad thriller does not leave one on the edge of the seat. Once these defining elements are removed, the movies appear almost identical, whatever one's initial spatial preferences might have been. As an analogy, suppose one restaurant critic enjoys seafood, while another enjoys pasta-based meals. Also suppose that both are served multiple dishes of each type that are heavily over-salted. We suspect that the original (latent) preferences will be non-observable, because the critics will dislike everything they receive. Here then, we suspect that the failure to select on (high) quality movies tends to disguise any spatial patterns in the data.

[Figure 6 about here.]

In Figure 6 we attempt to ameliorate this problem by presenting only those movies (with at least 15 reviews) that are 'high' quality. For present purposes this refers to those films that received a δ_1 score above the 80th percentile of all values of δ_1. In the figure, we also denote the (first) genre description of the movie as provided by *Rotten Tomatoes*, using different colors and plotting characters.

We now note several patterns that were unapparent before. First, movies of a similar genre appear in groups, running broadly north-west to south-east across the plot. In particular, in the right, bottom corner, foreign films (open triangles) cluster. North west of these come the dramas (filled circles). Running in a north-south band to the west of the dramas are the comedies, interspersed with the action/adventure pictures. The science-fiction fantasy movies (filled diamonds) appear to the west of the other movie types. In general,

drama movies score relatively highly on δ_3 (and this is also true of foreign films), and have higher δ_2 values also. By contrast, science-fiction fantasy films are low on δ_2 while comedies are somewhere between the two. Comedies though, tend to have lower δ_3 scores. Action adventure movies are similar to comedies in this regard

To construct Figure 7, we took a different tack: here, the movies are colored and demarcated by their *Motion Picture Association of America* rating. As can be seen from the figure, the bulk of the ratings are either R, which denotes that any viewer under 17 years of age requires an accompanying parent or guardian, or PG-13 which denotes movies for which "Parents [are] Strongly Cautioned" and that might be inappropriate for children under 13 years of age. Broadly speaking, the R rated movies lie predominantly to the north and east of the PG and PG-13 movies which themselves run in a broad band from the west to the east and south of the graphic. As a result, the more family-friendly pictures tend to score lower on the δ_3 axis, and although they are somewhat similar regarding δ_2. The 'unrated' movies help confirm this idea: generally lying to the north and east of the PG and PG-13 films, they include *Born into Brothels* which deals with the realities of child prostitution and *Capturing the Friedmans* which is a documentary concerning a father and son charged with child abuse. Presumably, neither of these films is suitable for minors.

[Figure 7 about here.]

Based on our assessment of Figure 6 and Figure 7, we present a combined graphic with our interpretation of the dimensions in Figure 8.

[Figure 8 about here.]

We label the west of the graphic as 'nerds', denoting that movies in this area are popular among sci-fi fans. To the north-east of the plot, we denote the area as 'art-house' to capture the fact that movies in this zone of the graphic might appeal to fans of (possibly pretentious, 'deep' and emotional) 'art-house' style pictures: *The Dreamers*, *In the Bedroom* and *Spider*

all reside in this general direction. By contrast, to the south of the plot, we denote the area as 'jocks' and the movies here are predominantly action-adventure/comedy combinations: we think *Gladiator* and *Anger Management* would appeal to such fans. Overlayed on this plot are two descriptors that refer to the ratings of the movies: 'adult entertainment' refers (broadly) to films that receive at least an R rating, while 'family fun' refers to all other movies. Now that we have gone some way to establishing the dimensions of movie criticism, the next section analyzes the effects of these judgements on movie success.

5 The Effect of Movie Reviews

We believe that movie critics, via their reviews, have a perceptible effect on the success of movie performance. In this section we measure that performance as 'profit' which we define as the difference between (the log of) a film's gross in the United States and the (log of) a film's production budget. We used data obtained from *The Numbers* website `http://www.the-numbers.com/`. The general theoretical assumption is that that film-makers seek to maximize revenue minus costs. In the subsequent section, we will report our findings on the relationship between movie reviews and *opening revenues*.

In addition to the reviews which are operationalized via our estimated $\hat{\boldsymbol{\delta}}$, we have several other predictors to act as 'controls': `rating`, which is a dummy for the MPAA rating the movie received; `create`, which is a dummy denoting the creative type of the movie: 'Contemporary Fiction', 'Factual' and so on. We use a production type dummy (`prod.dum`) which includes categories like 'live action' or 'stop motion animation'; a genre dummy (`genre.dum`) which denotes the movie's primary genre, such as 'drama' or 'romance'. We also record the movie's *initial* release in terms of the number of screens it was shown at when opening (`init.theat`) and its 'maximum' release in terms of the total number of screens it showed on during its *entire* theater run (`max.theat`) as well as using a dummy (`holiday`) to account

for possible profit variation due to the film's opening falling on a holiday. By *including* these variables in the estimation, some of which are surely contributing to the rating δs, we provide a more stringent test of any hypothesized relationship between reviews and box office success; that is, we are attempting to convince the skeptical reader that the δ scores are not simply proxies for more easily available, and better theoretically justified predictors. We thus hope to partially rule out the possibility that spurious correlations are driving any association we see in practice.

In Table 4 (on the left hand side) we report OLS results for our first model that includes all movies for which (complete) data is available; since the coefficients and other details on the controls ar not of current interest, we drop them, though readers can contact us directly if they wish to view them.

[Table 4 about here.]

Interestingly, δ_1 is the only significant predictor for movie success. Recall that δ_1 is essentially movie quality, so a positive coefficient makes sense: the better the critics thought the movie was, the better it does at the box-office.

We were surprised to see that neither δ_2 (which we think is related to 'nerdiness') and δ_3 (which we think connotes 'jockness' and/or 'art-houseness') is significant. We suspected though, that NSFC critics are not to everyone's tastes: they might not reflect the 'general' intended audiences for all the films. We thus split our sample into two parts: 'wide-release' movies that (by our definition) showed on *at least* 600 screens at the peak of their theater run, and 'independent' films that showed on *less than* 600 screens. To clarify, note that the industry standard defines a 'wide-release' as any film receiving an initial release of at least 600 screens. Problematically, some studios might release films for an initially 'limited' number of theaters to either (a) ensure their movie is eligible for Academy Awards (which requires it be released in a particular time frame for a given year) or to (b) 'test the waters' for a movie that might do poorly. We wanted to avoid counting such films as 'independent'.

The second column of Table 4 reports the wide-release regression: in practice, δ_1 has an increased p-value, and is no longer a predictor at the same significance level as before. This makes some sense if we regard the NSFC critics as being particular indicative of niche appeal.

The third column of Table 4 confirms these ideas: we now see that all the components of the $\hat{\boldsymbol{\delta}}$ estimate are significant at conventional levels for independent movies. Interestingly, 'nerdiness' (a low δ_2 value) is associated with more profitable films, and in fact, the coefficient is larger than previously. Now too, δ_3 is a significant predictor, although we note that more 'jock' movies tend to do *better* at the box office (relative to 'art-house' movies).

Broadly speaking, our results imply that the NSFC critical reviews are either disproportionately influential in convincing independent movie fans, or disproportionately representative of them. Neither is particularly surprising: these critics are known for their expertise and presumably more 'refined' tastes (in the same sense that a restaurant critic will probably not recommend a fast food joint as his top choice), so we expect their views to resonate with more selective audiences.

5.1 Movie Reviews and Opening Weekend Revenues

Independent movies—those which have a relatively small theater circulation as defined above—typically spend much less on advertising their film product than large-scale 'Hollywood' wide-releases. In part, this is a necessary feature of low budgets. A consequence is that we expect wide-release 'blockbuster' pictures to have much larger 'opening weekends' than independent movies, as audiences flock to theaters to see the latest release having been influenced by heavy publicity campaigns. We might also anticipate a different relationship between movie reviews and this opening revenue.

We defined our dependent variable as (the log of) the revenue made by movies between their opening Thursday (we look only at movies which did indeed open on a Thursday)

and the following Sunday. Again, we had a battery of controls as described above. In the bottom portion of Table 4 we report the regression coefficients for the $\hat{\boldsymbol{\delta}}$ we estimated for the movies. As can be seen, the movie quality dimension (δ_1) is not a helpful predictor for opening weekends of 'blockbusters' (column 4), yet the 'jock' dimension (δ_3) appears to be statistically significant.

In the fifth column of Table 4 we look at the more narrowly released 'independent' movies. Notice from the table that, now, the movie quality predictor δ_1 is a significant predictor of opening revenue, but that the other two, more substantive dimensions, are not.

All in all, it seems that opening weekends are differently structured across movie types: independent audiences need to believe the movie is high quality, whereas those seeing wide-release pictures are much less concerned. In part, we suspect this is due to the independent producers inability to advertise and generate 'buzz' for the films before the first weekend of viewing: instead, they must rely on solid reviews and helpful word-of-mouth.

6 Discussion

This paper developed a new 'utility threshold model' for estimating item response parameters of interest for movie critics and the films they review. We argued that a three dimensional spatial model was most appropriate and that the most important dimension represented movie 'quality', for which, universally, 'more' is preferred to 'less'. We presented evidence that such movie reviews are predictors of the financial success of movies, and that this effect is particularly strong for independent films.

In some IRT applications, notably educational testing, it makes sense to think of subjects and items in the *same* one-dimensional space: a test question has a particular 'difficulty' and a test-taker has an 'ability' on the same measurement line. In *multi*-dimensional spatial models where individuals make a binary choice—such as 'ideal point estimation' in

legislatures—items and subjects cannot usually be placed in the same space. Such models typically have micro-foundations in which actors make pairwise comparisons between two available alternatives (say, the 'status quo' and a legislative proposal) and select their preferred option. This is clearly not the case for critics: they choose to recommend a movie or not, without any attendant 'default' outcome. In light of this, we designed an approach with hybrid qualities: critics and movies *can* be located in similar (multidimensional) spaces and we are able to estimate individual 'quality' thresholds for the critics.

There are several avenues for further research. Clearly, most consumer-advice critics operate in similar ways to our movie-reviewers: restaurants, books, paintings, exhibits and so on are 'experienced' and then a judgement passed. More broadly, most 'satisfaction survey'-type exercises in marketing would yield data amenable to such analysis. We note that our framework can easily be extended to the case where individuals report multiple levels of satisfaction by incorporating more than one utility threshold. This would allow applications of our estimator to Likert scale data. In contrast to approaches relying on principal component analysis and related techniques, our estimator will produce estimates of product characteristics and rater ideal points in the same multidimensional space. In political science, promising applications include legislative cosponsorship and approval voting. Both of these have been studied to some degree using existing scaling techniques (Talbert and Potoski, 2002; Laslier, 2005), but we believe our approach can improve on these results by differentiating between spatial dimensions and heterogeneity in utility thresholds (following our argument in Section 3.2), and by providing estimates of the locations of bills and legislators, and voters and candidates, in the same multidimensional space.

A Identification of the Utility Threshold Model

In this section we provide conditions that ensure that the utility threshold model is identified.

Proposition 1 *Suppose that* $\boldsymbol{\alpha}_c = \boldsymbol{e}_c$ *where* $\boldsymbol{e}_c$ *is a unit vector for* $c \in \{1, \ldots, D\}$ *and* $\boldsymbol{\alpha}_{D+1} = \mathbf{0}$ *and* $\boldsymbol{W}_0$ *is a symmetric and positive definite matrix. Suppose that* F *is strictly increasing, that the vectors* $\{\boldsymbol{\delta}_{m,0} - \boldsymbol{\delta}_{m',0}\}_{m,m'}$ *span* $\mathbb{R}^D$*, and for any* $\boldsymbol{\omega} \in \mathbb{R}^D$*,*

$$[(\boldsymbol{\delta}_{m,0} + \boldsymbol{\delta}_{m',0})'(\boldsymbol{W}_0\boldsymbol{W}^{-1}\boldsymbol{W}_0 - \boldsymbol{W}_0) + 2\boldsymbol{\omega}'\boldsymbol{W}^{-1}\boldsymbol{W}_0](\boldsymbol{\delta}_{m,0} - \boldsymbol{\delta}_{m',0}) = 0 \text{ for all } m, m' \quad (12)$$

holds if and only if $\boldsymbol{W} = \boldsymbol{W}_0$*. Then there does not exist a parameter vector* $(\boldsymbol{\alpha}, \bar{u}, \boldsymbol{\delta}, \boldsymbol{W})$ *for which* $(\boldsymbol{\alpha}, \bar{\boldsymbol{u}}, \boldsymbol{\delta}, \boldsymbol{W}) \neq (\boldsymbol{\alpha}_0, \bar{\boldsymbol{u}}_0, \boldsymbol{\delta}_0, \boldsymbol{W}_0)$ *with* $\boldsymbol{\alpha}_c = \boldsymbol{e}_c$ *for* $c = 1, \ldots, D$ *and* $\boldsymbol{\alpha}_{D+1} = 0$ *such that*

$$F(\bar{\boldsymbol{u}}_c + (\boldsymbol{\alpha}_c - \boldsymbol{\delta}_m)'\boldsymbol{W}(\boldsymbol{\alpha}_c - \boldsymbol{\delta}_m)) = F(\bar{\boldsymbol{u}}_{c,0} + (\boldsymbol{\alpha}_{c,0} - \boldsymbol{\delta}_{m,0})'\boldsymbol{W}_0(\boldsymbol{\alpha}_{c,0} - \boldsymbol{\delta}_{m,0})) \quad (13)$$

for all c, m *holds.*

The restrictiveness of (12) is not immediately apparent, but the one-dimensional case is instructive. When $D = 1$, we have, $(\boldsymbol{\delta}_{m,0} + \boldsymbol{\delta}_{m',0})(\boldsymbol{W}_0 - \boldsymbol{W}) + 2\boldsymbol{\omega} = 0$ for m, m' such that $\boldsymbol{\delta}_{m,0} \neq \boldsymbol{\delta}_{m',0}$. If there are at least two distinct values of $\boldsymbol{\delta}_{m,0} + \boldsymbol{\delta}_{m',0}$, then it follows that $\boldsymbol{W}_0 = \boldsymbol{W}$ is the only possible solution to this system. Clearly, this is a very weak condition. In the multidimensional case, it is harder to reduce the condition in this way, but the condition is nonetheless likely to hold since we have a large number of equations ($DM(M+1)/2$) and very few free variables ($D(D+1)/2$).

Proof of Proposition 1: Consider any $(\boldsymbol{\alpha}, \bar{u}, \boldsymbol{\delta}, \boldsymbol{W})$ with $\boldsymbol{\alpha}_c = \boldsymbol{e}_c$ for $c \in \{1, \ldots, D\}$ and $\boldsymbol{\alpha}_{D+1} = \mathbf{0}$, where (13) holds. We show that such a point must satisfy $(\boldsymbol{\alpha}, \bar{u}, \boldsymbol{\delta}, \boldsymbol{W}) = (\boldsymbol{\alpha}_0, \bar{u}_0, \boldsymbol{\delta}_0, \boldsymbol{W}_0)$. Since F is strictly increasing, Equation (13) is equivalent to:

$$\bar{u}_c + (\boldsymbol{\alpha}_c - \boldsymbol{\delta}_m)'\boldsymbol{W}(\boldsymbol{\alpha}_c - \boldsymbol{\delta}_m) = \bar{u}_{c,0} + (\boldsymbol{\alpha}_{c,0} - \boldsymbol{\delta}_{m,0})'\boldsymbol{W}_0(\boldsymbol{\alpha}_{c,0} - \boldsymbol{\delta}_{m,0}) \ \forall \ c, m. \quad (14)$$

Factoring out (14), we obtain

$$\bar{u}_c+\boldsymbol{\alpha}_c'\boldsymbol{W}\boldsymbol{\alpha}_c-2\boldsymbol{\alpha}_c{}'\boldsymbol{W}\boldsymbol{\delta}_m+\boldsymbol{\delta}_m{}'\boldsymbol{W}\boldsymbol{\delta}_m = \bar{u}_{c,0}+\boldsymbol{\alpha}_{c,0}'\boldsymbol{W}_0\boldsymbol{\alpha}_{c,0}-2\boldsymbol{\alpha}_{c,0}'\boldsymbol{W}_0\boldsymbol{\delta}_{m,0}+\boldsymbol{\delta}_m'\boldsymbol{W}_0\boldsymbol{\delta}_{m,0} \ \ \forall \ \ c,m \tag{15}$$

$$\bar{u}_c+\boldsymbol{\alpha}_c'\boldsymbol{W}\boldsymbol{\alpha}_c-2\boldsymbol{\alpha}_c{}'\boldsymbol{W}\boldsymbol{\delta}_{m'}+\boldsymbol{\delta}_{m'}\boldsymbol{W}\boldsymbol{\delta}_{m'} = \bar{u}_{c,0}+\boldsymbol{\alpha}_{c,0}'\boldsymbol{W}_0\boldsymbol{\alpha}_{c,0}-2\boldsymbol{\alpha}_{c,0}'\boldsymbol{W}_0\boldsymbol{\delta}_{m',0}+\boldsymbol{\delta}_{m'}'\boldsymbol{W}_0\boldsymbol{\delta}_{m',0} \ \ \forall \ \ c,m. \tag{16}$$

Subtracting (16) from (15) yields, $\forall\, m, m'$,

$$\begin{aligned} -2\boldsymbol{\alpha}_c{}'\boldsymbol{W}\boldsymbol{\delta}_m \ + \ 2\boldsymbol{\alpha}_c{}'\boldsymbol{W}\boldsymbol{\delta}_{m'} + \boldsymbol{\delta}_m{}'\boldsymbol{W}\boldsymbol{\delta}_m - \boldsymbol{\delta}_{m'}'\boldsymbol{W}\boldsymbol{\delta}_{m'} \\ = \ \ -2\boldsymbol{\alpha}_{c\,',0}\boldsymbol{W}_0\boldsymbol{\delta}_{m,0} + 2\boldsymbol{\alpha}_{c,0}'\boldsymbol{W}_0\boldsymbol{\delta}_{m',0} + \boldsymbol{\delta}_{m,0}'\boldsymbol{W}_0, \boldsymbol{\delta}_{m,0} - \boldsymbol{\delta}_{m',0}'\boldsymbol{W}_0\boldsymbol{\delta}_{m',0}. \end{aligned} \tag{17}$$

When $c = D+1$, we obtain

$$\boldsymbol{\delta}_m'\boldsymbol{W}\boldsymbol{\delta}_m - \boldsymbol{\delta}_{m'}'\boldsymbol{W}\boldsymbol{\delta}_{m'} = \boldsymbol{\delta}_{m,0}'\boldsymbol{W}_0\boldsymbol{\delta}_{m,0} - \boldsymbol{\delta}_{m',0}'\boldsymbol{W}_0\boldsymbol{\delta}_{m',0} \ \ \forall \ \ m,m'. \tag{18}$$

Plugging (18) into (17), we obtain

$$\boldsymbol{\alpha}_c'\boldsymbol{W}(\boldsymbol{\delta}_m - \boldsymbol{\delta}_{m'}) = \boldsymbol{\alpha}_{c,0}'\boldsymbol{W}_0(\boldsymbol{\delta}_{m,0} - \boldsymbol{\delta}_{m',0}) \ \forall\ c,m,m'. \tag{19}$$

When $c \in \{1,\ldots,D\}$, Equation (19) yields

$$\boldsymbol{e}_c'\boldsymbol{W}(\boldsymbol{\delta}_{m'} - \boldsymbol{\delta}_{m'}) = \boldsymbol{e}_c'\boldsymbol{W}_0(\boldsymbol{\delta}_{m,0} - \boldsymbol{\delta}_{m',0}) \ \ \forall\, c \in \{1,\ldots,D\} \text{ and } m,m'. \tag{20}$$

Stacking these by column, we obtain

$$\boldsymbol{W}(\boldsymbol{\delta}_{m'} - \boldsymbol{\delta}_m) = \boldsymbol{W}_0(\boldsymbol{\delta}_{m,0} - \boldsymbol{\delta}_{m',0}) \ \forall\ m,m'. \tag{21}$$

Plugging Equation (21) into (19), we have

$$\boldsymbol{\alpha}_c'\boldsymbol{W}_0(\boldsymbol{\delta}_{m',0}-\boldsymbol{\delta}_{m,0})=\boldsymbol{\alpha}_{c,0}'\boldsymbol{W}_0(\boldsymbol{\delta}_{m,0}-\boldsymbol{\delta}_{m',0})\ \forall\ c,m,m'. \tag{22}$$

Since this must hold for all m, $\boldsymbol{W}_0$ has full rank, and the vectors $\{\boldsymbol{\delta}_{m,0}-\boldsymbol{\delta}_{m',0}\}_{m,m'}$ span $\mathbb{R}^D$ we have that

$$\boldsymbol{\alpha}_c=\boldsymbol{\alpha}_{c,0}\ \forall\ c. \tag{23}$$

Now plug Equation (23) into Equation (15) to obtain

$$\bar{u}_c+\boldsymbol{\alpha}_{c,0}'\boldsymbol{W}\boldsymbol{\alpha}_{c,0}-2\boldsymbol{\alpha}_{c,0}'\boldsymbol{W}\boldsymbol{\delta}_m+\boldsymbol{\delta}_m'\boldsymbol{W}\boldsymbol{\delta}_m=\bar{u}_{c,0}+\boldsymbol{\alpha}_{c,0}'\boldsymbol{W}_0\boldsymbol{\alpha}_{c,0}-2\boldsymbol{\alpha}_{c,0}'\boldsymbol{W}_0\boldsymbol{\delta}_{m,0}+\boldsymbol{\delta}_{m,0}'\boldsymbol{W}_0\boldsymbol{\delta}_{m,0}\ \forall\ c,m. \tag{24}$$

Using $c=D+1$ we obtain,

$$\bar{u}_{D+1}+\boldsymbol{\delta}_m'\boldsymbol{W}\boldsymbol{\delta}_m=\bar{u}_{D+1,0}+\boldsymbol{\delta}_{m,0}'\boldsymbol{W}_0\boldsymbol{\delta}_{m,0}\ \forall\ m\ . \tag{25}$$

We can subtract (25) from (24) to obtain

$$\bar{u}_c-\bar{u}_{D+1}+\boldsymbol{\alpha}_{c,0}'\boldsymbol{W}\boldsymbol{\alpha}_{c,0}-2\boldsymbol{\alpha}_{c,0}'\boldsymbol{W}\boldsymbol{\delta}_m=\bar{u}_{c,0}-\bar{u}_{D+1,0}+\boldsymbol{\alpha}_{c,0}'\boldsymbol{W}_0\boldsymbol{\alpha}_{c,0}-2\boldsymbol{\alpha}_{c,0}'\boldsymbol{W}_0\boldsymbol{\delta}_{m,0}\ \forall\ c,m. \tag{26}$$

When $c\in\{1,\dots D\}$, we obtain,

$$\boldsymbol{e}_c'\boldsymbol{W}\boldsymbol{\delta}_m=\frac{1}{2}(\bar{u}_c-\bar{u}_{c,0}-\bar{u}_{D+1}+\bar{u}_{D+1,0}+[\boldsymbol{W}]_{c,c}-[\boldsymbol{W}_0]_{c,c})+\boldsymbol{e}_c'\boldsymbol{W}_0\boldsymbol{\delta}_{m,0}\ \ \forall\ c\in\{1,\dots,D\}\text{ and }m,m', \tag{27}$$

where $[\boldsymbol{A}]_{i,j}$ denotes the element in the ith row of the jth column of the matrix $\boldsymbol{A}$. Stacking these by column, we obtain,

$$\boldsymbol{\delta}_m=\boldsymbol{W}^{-1}\boldsymbol{\omega}+\frac{1}{2}\boldsymbol{W}^{-1}\text{diag}\{\boldsymbol{W}\}-\frac{1}{2}\boldsymbol{W}^{-1}\text{diag}\{\boldsymbol{W}_0\}+\boldsymbol{W}^{-1}\boldsymbol{W}_0\boldsymbol{\delta}_{m,0}\ \forall\ m, \tag{28}$$

where,

$$\omega_c = \frac{1}{2}(\bar{u}_c - \bar{u}_{c,0} - \bar{u}_{D+1} + \bar{u}_{D+1,0}) \text{ for } c \in \{1, \ldots, D\}. \tag{29}$$

We can plug (28) into (18) to obtain

$$[(\boldsymbol{\delta}_{m,0} + \boldsymbol{\delta}_{m',0})'(\boldsymbol{W}_0\boldsymbol{W}^{-1}\boldsymbol{W}_0 - \boldsymbol{W}_0) + 2\omega'\boldsymbol{W}^{-1}\boldsymbol{W}_0](\boldsymbol{\delta}_{m,0} - \boldsymbol{\delta}_{m',0}) = 0 \; \forall \; m, m'. \tag{30}$$

By assumption, this is uniquely solved with,

$$\boldsymbol{W} = \boldsymbol{W}_0. \tag{31}$$

We can plug (31) into (30) to obtain,

$$\boldsymbol{\omega'(\delta_{m,0} - \delta_{m',0})} = 0 \;\; \forall \;\; m, m'. \tag{32}$$

Since this must hold for all m, m' and the vectors $\{\boldsymbol{\delta}_{m,0} - \boldsymbol{\delta}_{m',0}\}_{m,m'}$ span $\mathbb{R}^D$, we must have $\boldsymbol{\omega} = \mathbf{0}$. This implies that

$$\boldsymbol{\delta}_m = \boldsymbol{\delta}_{m,0} \; \forall \;\; m. \tag{33}$$

Plugging (31) and (33) into (24), we obtain $\bar{u}_c = \bar{u}_{c,0}$, thus proving the result. ■

References

Ainslie, Andrew, Xavier Drèze and Fred Zufryden. 2005. "Modeling Movie Life Cycles and Market Share." *Marketing Science* 24(3):508–17.

Albert, James H. 1992. "Bayesian Estimation of Normal Ogive Item Response Curves Using Gibbs Sampling." *Journal of Educational and Behavioral Statistics* 17(3):251–269.

Anand, Bharat N and Dmitri Byzalov. 2008. "Spatial Competition in Cable News: Where are Larry King and OReilly Located in Latent Attribute Space?" *Working Paper* Harvard University.

Andrich, D. 1988. "The Application of an Unfolding Model of the PIRT Type to the Measurement of Attitude." *Applied Psychological Measurement* 12:33–51.

Andrich, David and Guanzhong Luo. 1993. "A Hyperbolic Cosine Latent Trait Model for Unfolding Dichotomous Single-Stimulus Responses." *Applied Psychological Measurement* 17:253–276.

Beguin, A.A. and C.A.W. Glas. 2001. "MCMC estimation and some fit analysis of multidimensional IRT models." *Psychometrika* 66:471–488.

Blumer, Herbert. 1933. *Movies and Conduct*. New York: Macmillan.

Bock, R. D. and M. Aitken. 1981. "Marginal Maximum Likelihood Estimation of Item Parameters: An Application of the EM Algorithm." *Psychometrica* 46:443–459.

Bock, R and M Lieberman. 1970. "Fitting a response curve model for dichotomously scored items." *Psychometrika* 35:179–198.

Clinton, Joshua, Simon Jackman and Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98(2).

Coombs, Clyde. 1964. *A Theory of Data*. New York: Wiley.

DeSarbo, Wayne S. and Donna L. Hoffman. 1987. "Constructing MDS Joint Spaces from Binary Choice Data: A Multidimensional Unfolding Threshold Model for Marketing Research." *Journal of Marketing Research* 24:40–54.

Eliashberg, Jehoshua and Steven M. Shugan. 1997. "Film Critics: Influencers or Predictors?" *Journal of Marketing* 61(2):68–78.

Elsworthin, Catherine. 2005. "Sony to pay $1.5m for film hoax." *(Dublin) Independent* August 5.

Firth, David. 1993. "Bias reduction of maximum likelihood estimates." *Biometrika* 80:27–38.

Goettler, Ronald L. and Ron Shachar. 2001. "Spatial Competition in the Network Television Industry." *RAND Journal of Economics* 32:624–656.

Hambleton, Ronald K., H. Swaminathan and H. Jane Rogers. 1991. *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Press.

Heckman, James. 1981. In *Structural Analysis of Discrete Data With Econometric Applications*, ed. C. Manski and D. McFadden. Cambridge, MA: MIT Press chapter The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Time-Discrete Data Stochastic Process and Some Monte Carlo Evidence.

Hoijtink, H. 1990. "A Latent Trait Model for Dichotomous Choice Data." *Pychometrika* 55:641–656.

Hoijtink, H. 1991. "The measurement of latent traits by proximity items." *Applied Psychological Measurement* 15:153–170.

Hollinger, Hy. 2007. "MPA study: Brighter Picture for Movie Industry." *Hollywood Reporter* June 15.

Kamakura, Wagner A. and Rajendra K. Srivastava. 1986. "An Ideal-Point Probabilistic Choice Model for Heterogeneous Preferences." *Marketing Science* 5:199–218.

Kracauer, Stanley. 1957. *From Caligari to Hitler: A Psychological History of the German Film.* Princeton, NJ: Princeton University Press.

Laslier, Jean-Francois. 2005. "Spatial Approval Voting." *Political Analysis* 14(2):160–185.

Leenen, Iwin and Iven Van Mechelen. 2004. "A Conjunctive Parallelogram Model for Pick Any N Data." *Psychometrika* 69:401–420.

Lord, Frederic M. 1980. *Applications of Item Response Theory To Practical Testing Problems.* Mahwah NJ: Lawrence Erlbaum Associates.

Martin, Andrew and Kevin Quinn. 2001. "Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the US Supreme Court, 1953–1999." *Political Analysis* 10(2).

Maydeu-Olivares, Albert, Adolfo Hernandez and Roderick P. McDonald. 2006. "A Multidimensional Ideal Point Item Response Theory Model for Binary Data." *Multivariate Behavioral Research* 41:445–471.

Mulvey, Laura. 1975. "Visual Pleasure and Narrative Cinema." *Screen* 16(3):6–18.

Neelamegham, Ramya and Pradeep Chintagunta. 1999. "A Bayesian Model to Forecast New Product Performance in Domestic and International Markets." *Marketing Science* 18(2):115–136.

Ostini, Remo and Michael Nering. 2006. *Polytomous Item Response Theory Models (Quantitative Applications in the Social Sciences).* Thousand Oaks, CA: Sage Publications, Inc.

Poole, Keith. 2005. *Spatial Models of Parliamentary Voting.* Cambridge: Cambridge University Press.

Poole, Keith and Howard Rosenthal. 1991. "Patterns of Congressional Voting." *American Journal of Political Science* 35:228–278.

Poole, Keith and Howard Rosenthal. 1997. *Congress: A Political Economic History*. New York: Oxford University Press.

Rasch, Georg. 1961. *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Educational Research.

Riesman, David, Revel Denny and Nathan Glazer. 1968. *The Lonely Crowd*. New Haven, CT: Yale University Press.

Smith, Scott. 1998. *The Film 100: A Ranking of the Most Influential People in the History of the Movies*. Yucca Valley, CA: Citadel.

Takane, Yohsio. 1996. "An Item Response Model for Multidimensional Analysis of Multiple-Choice Data." *Behaviormetrika* 23:153–167.

Talbert, Jeffery C. and Matthew Potoski. 2002. "Setting the Legislative Agenda: The Dimensional Structure of Bill Cosponsoring and Floor Voting." *Journal of Politics* 64(3):864–891.

van der Linden, Wim J and Ronald K. Hambleton. 1997. Handbook of Modern Item Response Theory. New York: Springer chapter Item Response Theory: Brief History, Common Models, and Extensions.

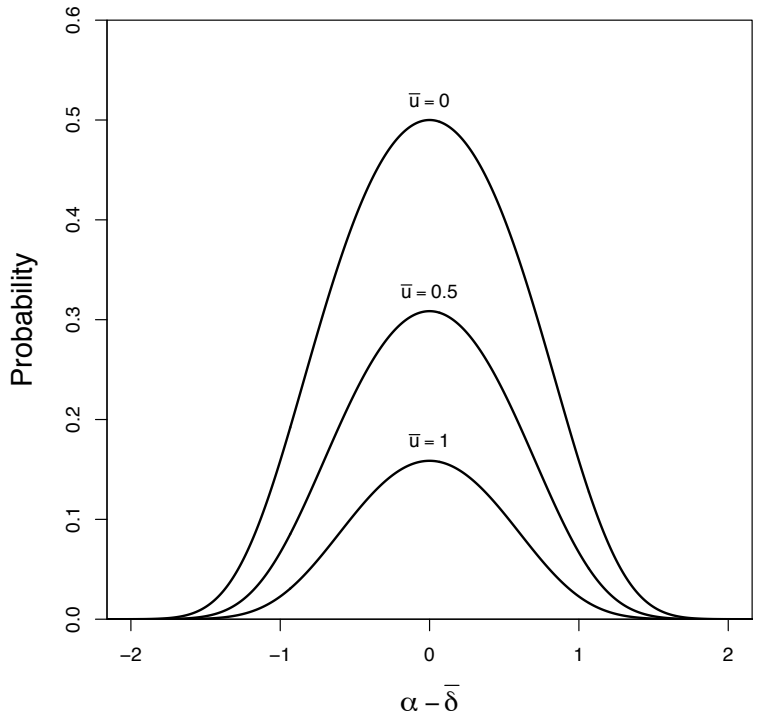

Figure 1: The 'trace line' from keeping the characteristics of the movie fixed (at $\bar{\delta}$) while (1) varying the spatial preference of the critic (α) and (2) varying the critic's utility threshold ($\bar{u}$).

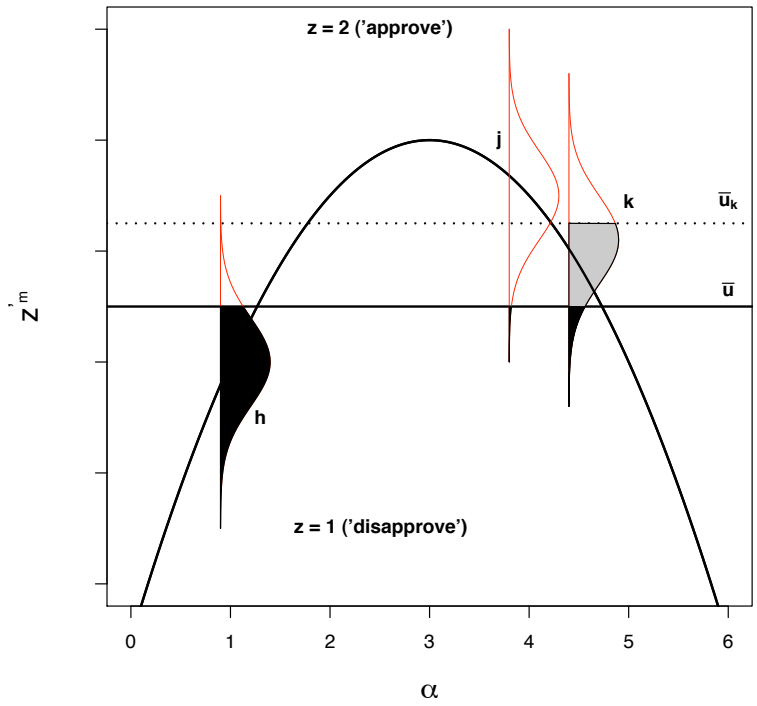

Figure 2: Critics with normal, homoscedastic error terms—and different spatial preferences (α)—contemplate the same movie: shaded areas correspond to disapproval.

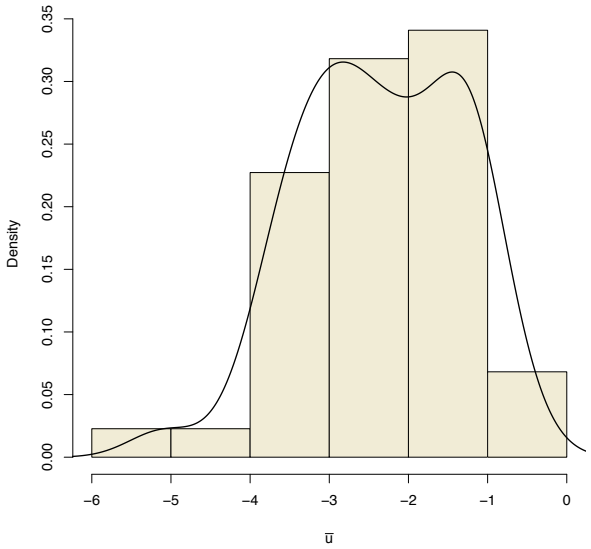

Figure 3: Density of estimated critic threshold utilities ($\bar{u}$)

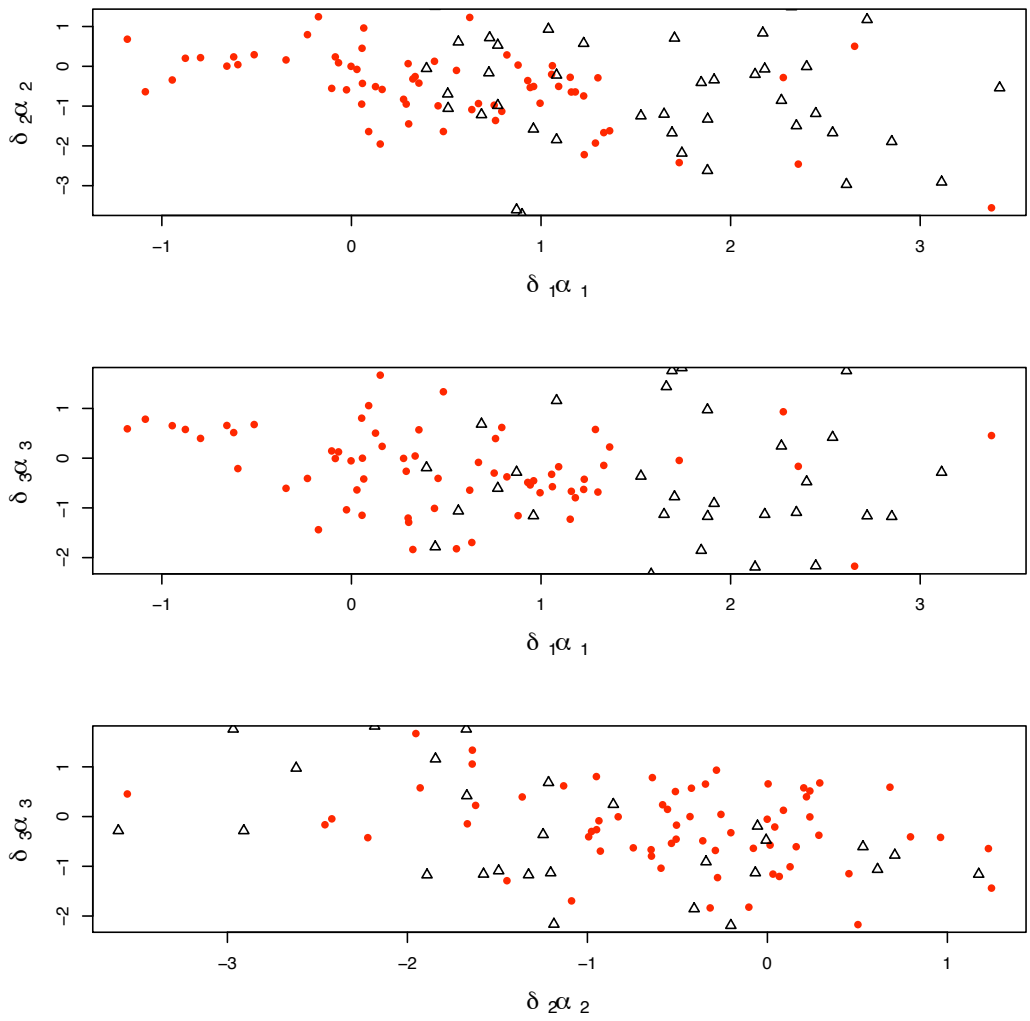

Figure 4: Scatter-plots for each of the three dimensions against the others. Movies are circular points, critics are dark triangles. Notice that the two groups show least overlap along the δ_1, α_1 axis.

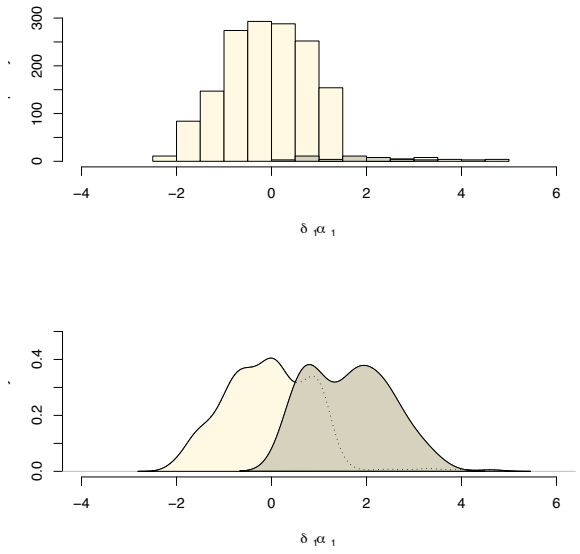

Figure 5: Histogram of movies (light color) and critics (dark color) in first dimension of model. We contend that this dimension is movie quality.

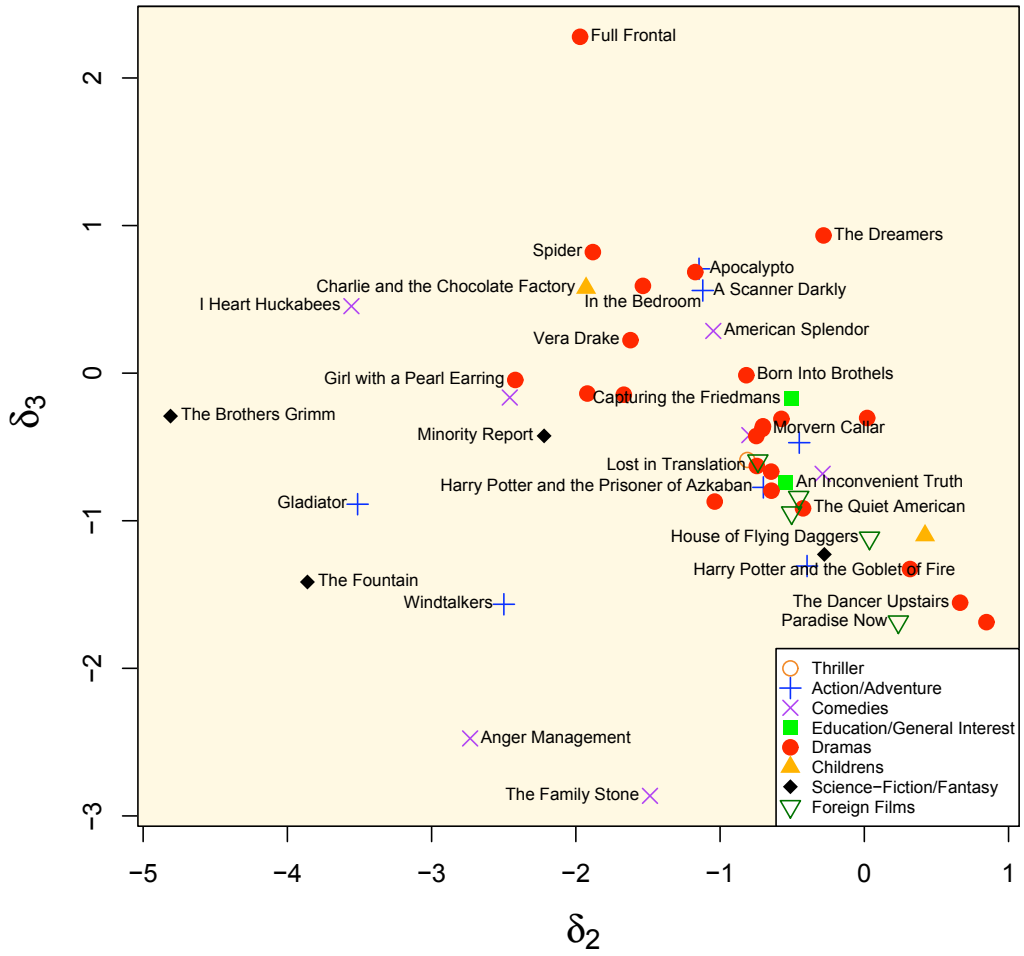

Figure 6: Scatterplot of movies in δ_2 and δ_3 space, plotting character and shade denote genres. Movies have 15 reviews or more, and are 'high quality'.

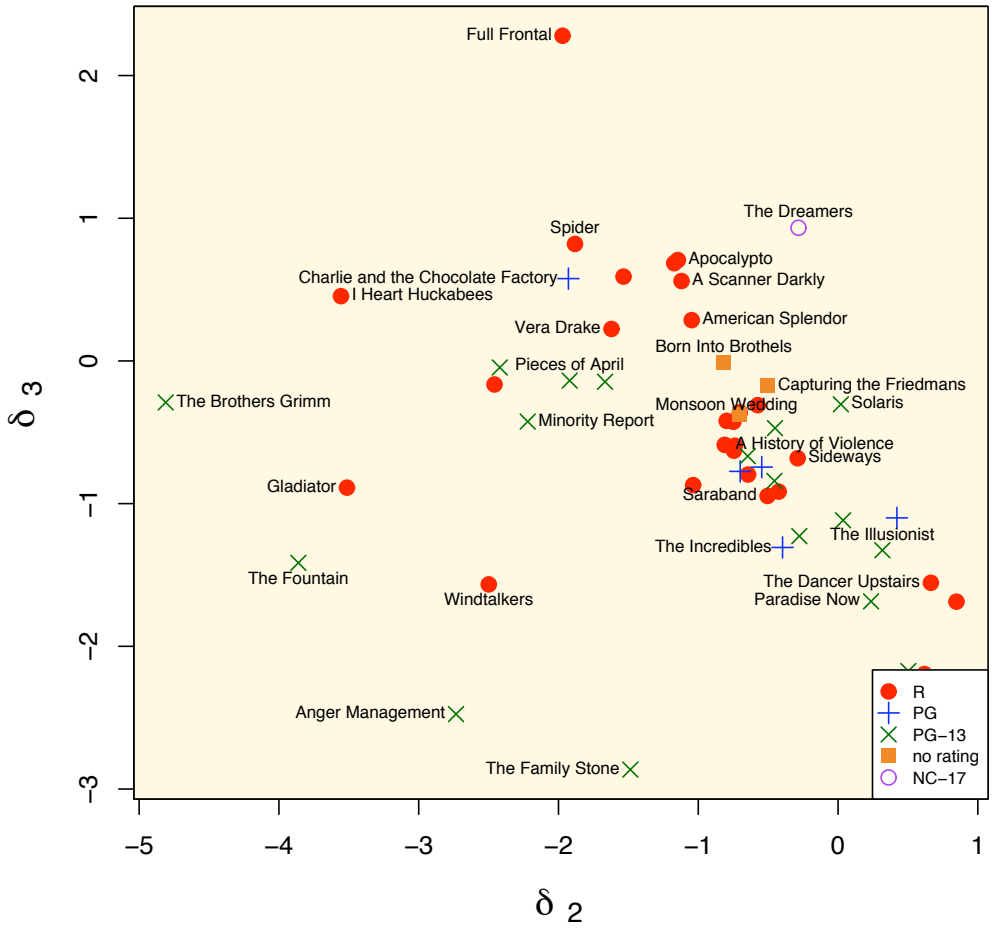

Figure 7: Scatterplot of movies in δ_2 and δ_3 space, plotting character and shade denote MPAA rating. Movies have 15 reviews or more, and are 'high quality'.

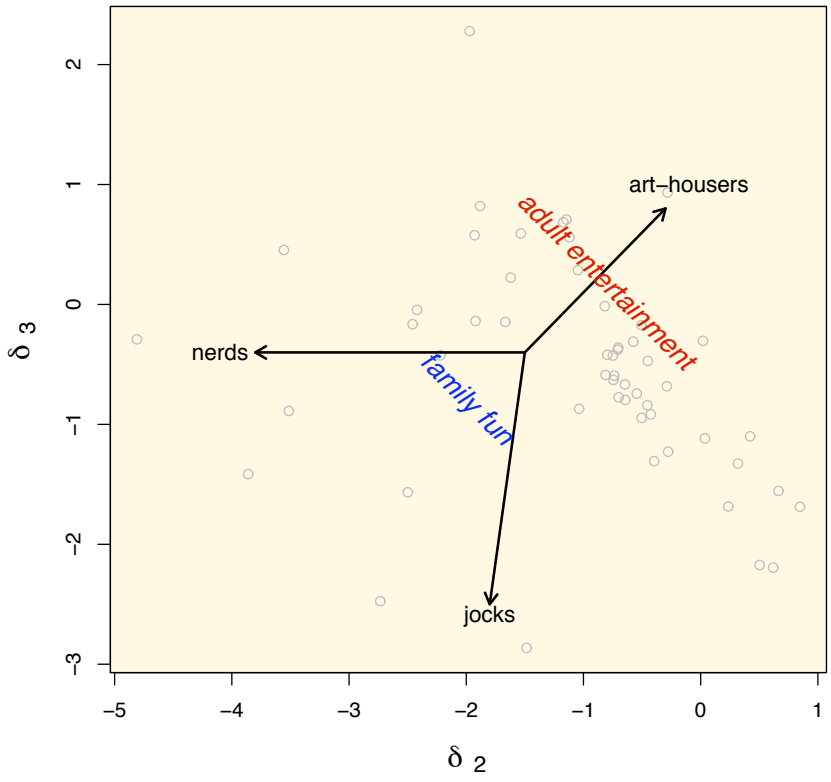

Figure 8: Scatterplot of movies in δ_2 and δ_3 space, with summary description. Movies have 15 reviews or more, and are 'high quality'.

	$D=0$	$D=1$	$D=2$	$D=3$	$D=4$	$D=5$	$D=6$	$D=7$	$D=8$
Geo Mean Prob (in sample)	53.0%	66.2%	71.1%	75.2%	79.1%	82.4%	84.7%	86.6%	87.9%
Geo Mean Prob (out of sample)	54.3%	63.8%	63.9%	64.6%	62.4%	64.1%	63.8%	63.3%	64.0%

Table 1: Goodness-of-fit statistics for each model (dimensions 0 through 8).

	Best Picture (AA)	Best Director (AA)	Best Drama (GG)	Best Director (GG)
δ_1	**0.672** [0.144]	**0.714** [0.141]	**0.645** [0.141]	**0.645** [0.152]

Table 2: Predicting 'Best Director' and 'Best Picture' Academy Award (AA) and Golden Globe (GG) winners and nominees with ordered probit. Predictor is δ_1 [standard error]. Emboldened coefficients are significant at $p < 0.01$ level.

Quantile	Title	Year	$\widehat{\delta_1}$	% 'fresh'	Genre
0.95	Lost in Translation	2003	1.23	95	Dramas
	Kontroll	2005	1.223	81	Foreign Films
	Primer	2004	1.22	72	Dramas
	The Last King of Scotland	2006	1.22	88	Dramas
	This Film is Not Yet Rated	2006	1.208	84	Comedy
	Captain Corelli's Mandolin	2001	−0.08	28	Dramas
	Blood Work	2002	−0.08	56	Dramas
	Veronica Guerin	2003	−0.08	52	Dramas
Median	Hearts in Atlantis	2001	−0.08	48	Dramas
	The Low Down	2001	−0.07	60	Comedies
	Birth	2004	−0.07	39	Dramas
	Juwanna Mann	2002	−1.57	9	Comedies
	Bulletproof Monk	2003	−1.58	22	Action/Adventure
	First Daughter	2004	−1.58	9	Comedies
	Jungle Book 2	2003	−1.58	20	Childrens
	Greenfingers	2001	−1.58	47	Dramas
0.05	Dragonfly	2002	−1.58	7	Dramas

Table 3: Movies at and around the 0.05, median and 0.95 quantiles of the empirical CDF of $\widehat{\delta_1}$. Final columns are *Rotten Tomatoes* aggregate rating and genre description from *Rotten Tomatoes*.

	Profit			**Opening Weekend**	
	All Movies Est[SE]	**'Wide release'** Est[SE]	**'Independent'** Est[SE]	**'Wide release'** Est[SE]	**'Independent'** Est[SE]
δ_1	**0.154** [0.057]	**0.179** [0.100]	**0.140** [0.071]	−0.058 [0.131]	**−0.271** [0.094]
δ_2	−0.057 [0.055]	0.042 [0.092]	−0.116 [0.069]	−0.088 [0.125]	−0.017 [0.091]
δ_3	−0.066 [0.052]	0.019 [0.086]	**−0.125** [0.066]	**0.263** [0.121]	0.081 [0.087]

Table 4: OLS results: top table are coefficients [Standard Errors] predicting profit (logged movie revenue minus logged movie cost). Dependent variable in right-side portion refers is opening weekend receipts. Emboldened coefficients are significant as $p < 0.10$ level

.